Health Data Analytics: Current Perspectives, Challenges, and Future Directions

Kavi Kumar Khedo, Shakuntala Baichoo, Soulakshmee Devi Nagowah, Leckraj Nagowah, Zahra Mungloo-Dilmohamud, Zarine Cadersaib, and Sudha Cheerkoot-Jalim

1 Introduction

The rise in data generation from electronic health systems is changing the landscape of healthcare as it is creating an immense demand for health data analytics, which aims at improving patients' care by deriving actionable insights from these datasets. The healthcare system worldwide is rapidly changing and adopting electronic health records. This is considerably increasing the quantity of clinical data that are available electronically. Moreover, an increasing number of mHealth tools are being adopted. Altogether, these are contributing to a rise in the generation of digital data. There is, therefore, a need for computational tools that can enable health practitioners to obtain new information from the massive data sets, known as medical big data. Immeasurable amounts of disparate medical data have become available in various healthcare institutions (consumers, providers, and pharmaceutical), ranging from patient records through medical imaging to clinical trials, among others.

Medical big data as compared to traditional big data are not very different. However, it is quite difficult to analyze medical big data; the size and intricacy of these datasets present great hurdles in analyses and consequent applications to a practical clinical environment. Additionally, medical data are difficult to access

K. K. Khedo · S. Baichoo · Z. Mungloo-Dilmohamud
Department of Digital Technologies, FoICDT, University of Mauritius, Réduit, Mauritius

S. D. Nagowah (✉) · L. Nagowah · Z. Cadersaib
Department of Software and Information Systems, FoICDT, University of Mauritius, Réduit, Mauritius
e-mail: s.ghurbhurrun@uom.ac.mu

S. Cheerkoot-Jalim
Department of Information and Communication Technologies, FoICDT, University of Mauritius, Réduit, Mauritius

N. Gupta, S. Paiva (eds.), *IoT and ICT for Healthcare Applications*, EAI/Springer Innovations in Communication and Computing,
https://doi.org/10.1007/978-3-030-42934-8_8

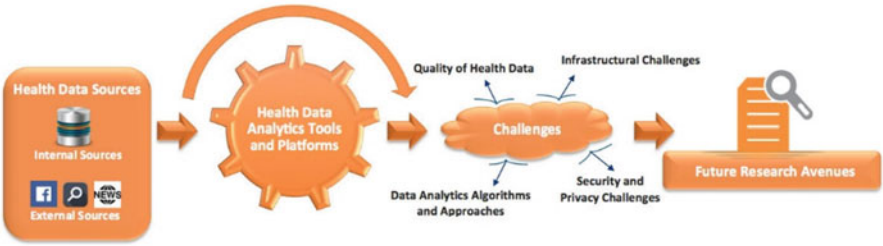

Fig. 1 Graphical summary

for the risks of misuse, as patient details are expected to remain confidential, unless patients' approval is sought. Despite these hurdles, big data technology is already being used in healthcare, for example, for predictive analytics [6] and to reduce healthcare costs [6]. Health data analytics utilizes numerous techniques, including data modeling, data mining, as well as machine learning, among others, to evaluate historical and real-time data in view of predicting future outcomes. It can be used to improve the overall management of healthcare systems from end to end.

This chapter provides an overview of the main application areas that can benefit from health data analytics, by providing practical examples and emphasizing on the type of data they use with their sources, as well as listing their benefits and challenges. A nonexhaustive list of tools and platforms used in the context of health data analytics is provided. Finally, a thorough discussion on the future and existing research opportunities in health data analytics is provided.

Methodology

In order to conduct the review of health data analytics in this chapter, the following methodology has been adopted.

1. The main application areas of data analytics in healthcare are discussed.
2. The different sources of health data are analyzed.
3. State-of-the-art health data analytics tools and platforms (based on popularity and adoption rate) are surveyed, reviewed, and compared.
4. Use cases for the application of health data analytics are described.
5. The main challenges for the use of health data analytics in healthcare institutions and possible solutions are discussed.
6. Future directions for health data analytics are elaborated.

Figure 1 provides a graphical summary of the chapter.

2 Application Areas for Health Data Analytics

Health data analytics are progressively being used by different healthcare institutions to improve effectiveness of the services being provided. This section describes the main application areas of data analytics in healthcare. A summary of these

application areas is provided at the end of this section, which highlights the sources of data, benefits, and challenges.

2.1 Data Analytics for Drug–Disease Association

Currently, over 23 million published biomedical research articles, clinical case reports, and randomized controlled trials are available on MEDLINE, a US bibliographic database in life sciences, mostly geared toward biomedicine [60]. MEDLINE, therefore, contains biomedical information, including drug–disease associations, from various sources. However, most of the available knowledge is in textual format and has limited machine understandability. Several works, such as Chen et al. [15], Chiang and Butte [16], Frijters et al. [27], and Xu and Wang [88], have focused on different text mining techniques to extract co-occurring concepts, particularly drug–disease treatment associations, from the available biomedical literature. Many of the studies involved the use of knowledge sources such as UMLS (Unified Medical Language System) Metathesaurus and the comparison with established clinical knowledge sources, such as ClinicalTrials.gov [16, 88] and DRUGDEX [16].

2.2 Data Analytics for Disease Outbreak Detection and Surveillance

A number of disease outbreak detection and surveillance systems have been developed to help in the early detection of outbreaks and to assist in outbreak investigation and propagation. These systems heavily use data analytics. Lopez et al. [53] proposed a framework to predict H1N1 influenza epidemic based on H1N1 prevalence and climatic data accumulated from previous years in India. Sources of data include huge repositories of health and geospatial data. By applying analytics to these huge sources of data, risk factors and geospatial vulnerabilities have been determined. The result of these analytics can help in developing control and preventive strategies for influenza epidemics, allowing effective use of limited resources in the public health sector. Another model is proposed by Culotta [18], where the author analyzes messages posted on a social website (Twitter) to determine the correlation between search term frequency with influenza statistics with the aim of responding rapidly to health epidemic. Twitter has been chosen mainly because of its diversity of users. Five hundred thousand Twitter messages were collected from a 10-week period. Simple linear regression and multiple linear regression models were used to predict influenza-like illness (ILI) rates based on the frequency of messages related to specific keywords.

2.3 Data Analytics for Pharmacovigilance

Adverse drug reactions (ADRs) are those that are generally caused by the intake of medicines. Pharmacovigilance refers to all the activities that relate to the detection, assessment, understanding, and prevention of adverse effects associated with prescribed drugs [76]. Nikfarjam et al. [64] designed a machine-learning-based approach to extract indications of adverse drug reactions from the unstructured text in social media. User posts were collected from two different social media sites, mainly Twitter and DailyStrength. Two experts then independently annotated the posts as being an adverse drug reaction, an outcome of interest, an indication, and any other mention of symptoms. A supervised sequence labeling conditional random field (CRF) classifier was then used to extract the adverse drug reaction concepts from user posts.

Pharmacovigilance programs mostly rely on spontaneous reporting. The US Food and Drug Administration (FDA) adverse event reporting system (AERS) is such an example. There is, however, a need to exploit other nontraditional resources generated by patients on the Internet. These include social media and the logs of popular search engines, which might contain useful health information. White et al. [87] used such search logs for systematic signal reporting and also assessed the potential of using these search logs for the generation of early warnings about adverse drug reactions. On another note, Harpaz et al. [33] stated that due to certain limitations, the early stage of development, and the complexity of adverse drug reactions, it is still not possible to make clear-cut statements about the ultimate utility of the concept.

2.4 Data Analytics for Healthcare Management

Everyday healthcare institutions around the world capture information electronically ranging from hospital data to patient data. The amount of data generated is huge, and data analytics can play a pivotal part in enhancing human health and reducing costs in healthcare institutions [86]. One main example of application of data analytics in healthcare management is monitoring consequences like infections at surgical sites, rate of readmission of patients, patient waiting times, and room utilization in healthcare institutions. Another example is the identification of patients and staff [42, 90] across the hospital using radiofrequency identification (RFID). The data captured using RFID can be used to model patient flow in hospitals as well as care delivery processes and to monitor compliance with hospital policies [68]. For disease management, Moore [62] has proposed the use of data analytics in the monitoring of patients with chronic diseases like diabetes and hypertension, for reaching specific blood sugar and blood pressure targets using smart devices embedded with RFID which communicate with their providers. Other example applications include the monitoring of elderly patients to prevent falls [72].

2.5 Data Analytics for Clinical/Medical Research

Diseases evolve over years and the same drug cannot effectively combat a mutated form of the same disease. In the healthcare and pharmaceutical industries, there is a surge in structured and unstructured biomedical data generation from several sources, including the output from surveys and research experiments gathered by pharmaceutical companies, patients, healthcare providers, and social media [14]. In order to make good use of the data, it is very important to integrate all these heterogeneous data and devise smart algorithms to link them. Integration of disparate data can help in identifying and establishing the worthiness of new drug targets, support early identification of safety and efficacy issues, and improve patients' treatment.

Antibiotics are normally used to treat bacterial infections. However, there is increasingly a rise in resistance to antibiotics that once cured ailments across the spectrum, and this is turning into a potential source of prolonged illness known as "super-bugs," which can lead to disability and death. Antibiotic resistance (ABR) occurs either due to natural mutations of diseases or because of unregulated and frequent use of antibiotics, whereby the pathogens develop resistance against a specific antibiotic. Big data analytics can also be helpful in managing the problem of ABRs and in tracking the production to the distribution of antibiotics in the retail market. The data generated can also be used for developing statistical models to show the relationship between antibiotic consumption and associated resistance.

2.6 Data Analytics for Clinical Practice

Cancer is a very complex disease with only one tumor having billions of cells each capable of mutating. The genetic makeup of the cells needs to be captured to understand how they evolve and adapt. These data need to be acquired often to provide a clear picture of the disease, and hence, huge quantities of data are generated. The volume of published cancer research has increased significantly along with a rapid increase in patient-specific information such as genomic data. Clinical trials work by testing new treatments in small cohorts at initially, looking at how well the treatment works and to identify any adverse effects. If the results of a trial prove encouraging, it is expanded to include larger groups of people. However, in most of the cases, enrolled patients do not represent the general population of patients with cancer. Moreover, real-world patients with cancer tend to be older, sicker, and more ethnically diverse than the typical study patients who tend to be in good health – except for the cancer. CancerLinQ project (https://cancerlinq.org/) by American Society of Clinical Oncology (ASCO) aims at providing oncologists with data from a large number of patients, with lifelong learning opportunities and a common platform for all cancer stakeholders. CancerLinQ is making unparalleled use of massive amounts of medical records of patients with cancer to discover

patterns and trends and to measure their care against that of similar products and recommended guidelines.

2.7 Summary of Application Areas for Health Data Analytics

In this section, a summary of the six application areas for health data analytics is given. The type of data analytics used, the sources of data, the benefits and challenges for each application area are highlighted in Table 1.

As it can be observed from Table 1, there is a wide range of data analytics techniques being used in the healthcare sector ranging from simple regression methods to sophisticated machine-learning algorithms. One of the main sources of data remains the electronic patient records as it contains valuable information about the diagnosis, disease, and treatment. However, it is noted that there is a growing number of online open databases for diseases, drugs, and genes that are becoming important sources of data for health data analytics. Even social media such as Facebook and Twitter are becoming important sources for health data. In all the application areas investigated, it is clear that health data analytics are increasingly being used and several benefits are being derived, such as the discovery of new knowledge, improving quality of services, early detection of disease outbreak, new drug development, and cost-effective operations. However, there are still a number of challenges to overcome in the area of health data analytics. The reliability and accuracy of the health data analytics need to be improved, and there are integration issues to be handled for these large volumes of data emanating from diverging sources.

3 Health Data Sources and Types

Health data used for analytics can be obtained from various sources as presented in Sect. 2. These sources can be categorized into internal sources (such as data from electronic patient records) and external sources (such as online open databases) [56]. Each of the sources contains different data types with varied level of details. Moreover, the reliability, accuracy, and accessibility of the data vary. In this section, the different sources of health data, categorized as internal and external, together with their characteristics, are detailed.

3.1 Internal Data Sources

Healthcare providers routinely produce a large amount of data as part of their normal operations. Since these data are produced from internal data management systems,

Table 1 Summary of application areas for health data analytics

Application area	Type of data analytics	Sources of data	Examples	Benefits	Challenges
Drug–disease association	Sequence, structure, pathway, text analysis, pattern learning, guilt-by-association, and statistical techniques	DrugBank database, CORUM database, FunDO database for diseases, PPI network data, MEDLINE, and DRUGDEX reference standard	Drug repurposing, novel drug use suggestions	Application of known drugs to treat new diseases (better than new costly drug discovery which is lengthy and risky)	Reliability and accuracy of the data analytics techniques
Disease outbreak detection and surveillance	Regression analysis, multiple linear regression models, keyword matching, and classifier	Geospatial databases, health records, climatic bulletins, social media	Prediction of H1N1 influenza epidemic, prediction of influenza-like disease	Effective prevention and control strategies, optimized resource allocation	Prediction model affected by other factors (e.g., biological and sociobehavioral attributes)
Pharmacovigilance	Supervised sequence labeling CRF classifier, self-controlled study designs	DailyStrength, Twitter, Internet search logs	Extract complex medical concepts with relatively high performance, ADR detection	Analysis of highly informal text in social media, ADR detection	Misspellings, abbreviations, and phrase construction irregularities
Healthcare management	Text analysis techniques, pattern matching, empirical/statistical analyses, and data mining techniques	Healthcare institutions records (patients, rooms allocation, surgical instruments RFID)	Optimal allocation of surgical instruments and rooms to patients	Better service provision, reduce waiting time for patients, and cost-effective operations	Large amount of data to analyze and multiple systems
Clinical/medical research	Predictive analytics, statistical modeling, survival analysis, genomic analytics and machine learning (including images)	Sequencing and gene expression data, drug data, protein and drug interaction data, clinical trial, medical images, and electronic patient record data	Diagnostic of diseases based on image analysis, prediction of new potential drugs, drug repositioning	Fast development of new drugs, determining most appropriate treatments using medical images	Integration of disparate sources of data, need of complex data mining algorithms specially for medical images
Clinical practice	Machine learning	Patient data, prescription data, data related to observational studies	Uncover patterns and trends in medical records of patients, connects and analyzes real-world data from different sources	Improving medication adherence and patient care	Mining heterogeneous information from different sources

they are considered as internal data sources. Although these data are collected routinely, they are seldom collected with data analysis in mind and, hence, need to be processed to be useful [82]. Data from internal sources often reside on separate information systems (ISs), and therefore, correlations between those data cannot be made and useful insights cannot be derived. Table 2 lists the various internal data sources identified, their data types, and characteristics.

Internal data are obtained from disparate information systems. The data are normally structured and reliable. Internal data contain mainly individual information about patients and information about internal processes of the healthcare institutions. Internal data are normally obtained routinely, and therefore, over time, there is a large volume of data which are regularly archived. Data analytics may be applied to internal sources in order to gain insights into diseases trends, effectiveness of treatments, cost-effectiveness of operations, proper management of resources, and performance of individual departments.

Table 2 Internal data sources

Internal data sources	Types of data	Characteristics of the data
Clinical data	Medical imaging, diagnosis, physician advice, and notes and medication nonadherence	Data routinely collected. No exact standard or format for storing physician notes or medical imaging. Medical imaging not always part of main IS (e.g., separate electronic files)
Electronic patient records (EPRs)/electronic medical records (EMRs)	Patient details, medical history, appointments, diagnosis, treatment details, and prescription requests	Patient data are routinely updated. Main source of patient information. Data stored using common medical record standards
Health monitoring data	Heartbeat, blood pressure, food intake, daily walking distance, and activities	Fine-grained individual data. Often not recorded on ISs. Volume of data is huge
Pharmacy information system	Data specific to drug prescription (e.g., patient name, ID, date, time and duration, drug dosage, administration route, frequency of intake), drug stock information, and expiry dates	Routinely updated. Data stored using common formats. Usually a separate IS from main healthcare provider system
Laboratory information management system (LIMS)	Data related to laboratory samples and results of different tests	Data usually stored in a separate IS from the healthcare provider's main system. Data are of technical nature and require expert knowledge for interpretation
Real-time location tracking systems	Location of assets and medical equipment, staff, and patients	Location information is collected routinely. Information collected from localization technologies (e.g., RFID, beacons, tags, and sensors)

3.2 External Data Sources

While internal sources of data reflect those data that are under the control of the healthcare institutions and are heavily regulated, external data are not generated in the healthcare environment. However, external data supplement internal data and can give useful insights into health-related conditions as well as help into making important decisions in this sector. Table 3 provides a nonexhaustive list of external data sources. The data types and characteristics are also described.

Increasingly, there is a huge amount of data that is being created at an unprecedented rate from external sources such as the Internet. There are a number of open online databases, from which very useful and reliable health data can be obtained. While data from internal sources are highly localized data from individual healthcare institutions, external sources mainly provide data from all over the world. However, data from external sources are mainly unstructured and are normally not stored using common standards and formats. Moreover, accuracy of the data cannot be guaranteed in most of the cases. Data analytics on external data sources can provide very important insights into disease outbreaks around the world, behavioral patterns of patients, and health events in different countries.

4 Health Data Analytics Tools and Platforms

As already elaborated in the previous sections, healthcare institutions are generating large volumes of data from various sources. Therefore, these healthcare organizations are seeking to harness the potential of their data. In recent years, numerous health data analytics tools and platforms have been developed. The capabilities of these tools and platforms are also evolving rapidly. They are increasingly being used for drug discovery, better decision support, predictive analysis, and evidence-based treatment. This section explores different tools and platforms (including a comparative analysis), types of data manipulated by these tools, data analytics techniques used, and the trend of these tools.

4.1 Tools and Platforms

In this section, state-of-the-art health data analytics tools and platforms are analyzed. These tools and platforms have been selected based on their popularity and adoption rate.

Table 3 External data sources

External data sources	Type of data	Characteristics of the data
Biomedical literature	Journals, online databases, books, proceedings, newsletters, technical reports, web pages, citation database for peer-reviewed literature from life sciences, health sciences, and so on	Embase by Elsevier is subscription based. Medline by US National Library of Medicine (NIH) is open access. Both are regularly updated. Large volume of data from multiple sources
Clinical Trials (ClinicalTrials.gov)	Experiments and observations done in clinical research	ClinicalTrials.gov (US) database: publicly and privately supported clinical studies conducted on participants from many countries, open access, and regularly updated
DRUGDEX	Reference standard for drug use review	Covers FDA-approved and investigational prescription and nonprescription drugs, as well as non-US preparations. DRUGDEX updated weekly and in each Metathesaurus release as part of RxNorm (standardized nomenclature for clinical drugs for humans, produced by US National Library of Medicine)
Spontaneous reporting systems (SRSs). E.g., US Food and Drug Administration (FDA) Adverse Event reporting System (AERS)	Reports of suspected adverse drug reactions (ADRs) that typically capture suspected and concomitant drugs, indications, suspected events, and limited demographic information	Raw data consisting of individual case safety reports extracted from the FAERS database are available to public through creation of relational databases. AERS has millions of drug–event combinations.
Health social networking sites (SERMO, PatientsLikeMe, and MedHelp)	Patients' profiles, health conditions, medications. PatientsLikeMe has an openness policy (users can agree to share all their health data). Collects and stores two types of data from users: shared data (e.g., biography, conditions, treatments, symptoms, outcomes, laboratory results) and restricted data	Data are mostly unstructured from dispersed geographical areas. Large volume of data posted daily. Accuracy of data cannot be guaranteed
Generic social networks such as Twitter	Individual posts about health, medications, doctors' visits	Data are unstructured. Requires data mining and natural language processing for capturing health-related information
Users' web search logs	Time stamp information, location of user, disease information, drug information, behavioral data, health-related symptoms, and frequency of search (to indicate seriousness of health condition)	Can be categorized into two main types, namely: search logs (user raises query to a search engine and selectively clicks on the search results returned) and browse logs (user visits a web page other than a search result page)
Large repositories of geospatial and health data, e.g., HealthMap	Latest real-time disease outbreak information, H1N1 prevalence, and climatic data accumulated from previous years	Real-time data from large geographical areas. Reliable disease outbreak information. Regularly updated.

4.1.1 Philips HealthSuite Digital Platform

Philips HealthSuite Digital Platform [37] is an open platform available on the cloud, which aims at collecting, integrating, and analyzing clinical and other data from different internal and external sources, including sensors and devices. It uses Amazon Web Services (AWS) to securely store about 15 PB of patient data that can be analyzed to give valuable insights to patients and healthcare specialists. The platform uses contextual predictive analytics, decision support algorithms, and machine learning to analyze data to provide useful insights on patients' health.

4.1.2 EMIF (European Medical Information Framework) Platform

The European Medical Information Framework [24] Platform is a common information platform that links up clinical information of about 50 million patients around Europe. It integrates and harmonizes data from various data sources, such as population-based registries, national registries, biobanks, and hospital-based registries, to allow reuse of data. Additionally, the platform makes data available for browsing and exploitation by the user, analyses data, and provides support for data visualization. A number of techniques such as Spearman correlation analysis, hierarchical clustering, K-means clustering, principal component analysis, linear regression, survival analysis, and statistical methods have been used by the platform to analyze data.

4.1.3 Hortonworks

Hortonworks [36] is bringing a revolution, which aims at transforming big data analytics in healthcare and medicine by making healthcare data available and accessible. It offers open-source connected data platforms, which are based on Apache Hadoop and Apache NiFi. The potential benefits of Hortonworks Data Platform (HDP) include the use of accumulated information over time for healthcare predictive analytics with feeding algorithms predicting the probability of an emergency earlier than it could be detected with traditional bedside visits.

4.1.4 IBM Analytics

IBM Analytics (https://www.ibm.com/analytics/us/en/) solutions enable healthcare organizations to connect information from disparate sources into a single-trusted view with the eventual delivery of insights into the data. Big data analytics are mainly possible due to the accuracy and performance of the IBM InfoSphere Master Data Management. The Watson Health Cloud collates large amounts of medical data into a centralized cloud-based repository. The IBM platform brings advanced

analytics, by reading 200 million pages of text in 3 s, which help turn the raw and disparate data into vital insights that are beneficial to the patients.

4.1.5 Sagitec HealHub Digital DataOps Platform for Healthcare and Life Sciences

HealHub [74] is a DataOps platform for healthcare organizations. It enables real-time collaboration between data analysts and the teams involved in the day-to-day running of the hospital. It restructures quality information from disparate sources to provide a better understanding and for increased revenue. When data are ingested into the platform, the information is automatically cleaned and scrutinized according to protocols, standards, and master data. The system accepts data in all formats. It can capture data from health and medical devices, which is then integrated into the system. In addition to sharing operational data with internal analysts, HealHub also integrates with external systems, such as customer relationship management, clinical data management, and data warehouses; with online libraries and websites like PubMed and ClinicalTrials.gov; with ratings from doctors; with reviews from patients and Facebook/Twitter; and with other data channels. It provides visualization capabilities as well as what-if analysis.

4.1.6 Sisense

Sisense Business Intelligence [35] software provides hospitals and other healthcare organizations with a robust business analytics solution to structure, analyze, and visualize complex data. It supports healthcare organizations in providing enhanced services to patients or clients by closely monitoring various metrics and key performance indicators (KPIs), with timely and accurate data which can be medical, financial, and administrative. The dashboards in the software provide detailed information to track individual performance. Sisense has robust Extract, Transform and Load (ETL) and data modeling capabilities, advanced analytics and statistics, and an intuitive user interface for creating dashboards, generating reports, and providing visualizations.

4.1.7 EHDViz: Clinical Dashboard Development Using Open-Source Technologies

EHDViz [5] is a clinical dashboard development framework toolkit for generating web-based, real-time clinical dashboards for visualizing dissimilar biomedical, healthcare, and well-being data. It is developed as an extensible toolkit that uses R packages for data management, normalization, and producing high-quality visual

representations over the web using R/Shiny web server architecture. EHDViz can integrate data from different sources, such as biomedical and healthcare data visualization for a combined health assessment. EHDViz can also be used as a toolkit to emulate EHR environment to increase simulation-based learning. Hospitals and healthcare systems are emerging as learning health systems (LHS), and thus, data capture, smart clinical dashboards, and adjustive visual analytics can play an important role in managing the patient population.

4.1.8 ICDA: A Platform for Intelligent Care Delivery Analytics

Intelligent Care Delivery Analytics platform (ICDA) [29] conducts risk assessment analytics by processing large collections of ever-changing electronic medical data to identify high-risk patients, in view of improving patient outcomes and managing costs. ICDA works by integrating large volumes of data into an integrated data model, then effecting a collection of analytics that identify at-risk patients. It also provides an interactional environment through which users can access and critically evaluate the analytics results. Additionally, ICDA provides APIs via which analytics results can be extracted to interface in external applications.

4.1.9 EMC Healthcare Analytics

Dell EMC Services provides healthcare institutions with the capabilities to assess, prove, and deploy data analytics use cases, as well as the technology, capabilities, and platform required to support them. The EMC Healthcare Analytics Solution [23] allows the aggregation, analysis, and visualization of data to make informed patient care and operational decisions. Dell [19] describes some sample use cases such as reduction of hospital readmission and improving patient care.

4.1.10 OpenML Platform

Open Machine Learning is a collaboration platform, which is designed to organize datasets, machine-learning workflows, models, and their evaluations [84]. Scientists and researchers can share machine-learning data sets, code, and experiments for more effective, large-scale, real-time collaboration. Its easy-to-use APIs help to automate many tedious research tasks. It is also easily integrated into several machine-learning tools. Researchers can share their latest results, thus speeding up research through combined and linked results. OpenML can be accessed through some interfaces, like R and WEKA. Some researchers have proposed extensions of the OpenML platform in applications like healthcare predictive analytics, which allow secure sharing of data, workflows, and evaluations.

4.1.11 Generic Tools and Platforms: Google Health, Apple HealthBook, and Microsoft HealthVault

Information captured through new types of sensors and systems is integrated in databases facilitating data mining and revealing new insights. Several companies have created health toolkits, such as Google Health [81], Google Fit, Samsung S.A.M.I, Microsoft HealthVault, and Apple HealthBook [43]. Google Health can be used to automatically import health records, test results, and prescription history. Other services include appointment scheduling, prescription refills, as well as wellness tools. Microsoft HealthVault, also known as the "PayPal for health information," consists of two main products, namely an electronic repository for health data and a specialized search engine for health information. Comparing Microsoft HealthVault and Google Health showed that they were both extendible platforms where functionalities vary on configurations used as well as affordability toward services. No deductions could be made as to which of these two systems are better in terms of functionality. Besides, Microsoft HealthVault can only be used in the United States [81].

According to Schulz and Neumann [77], applications such as Apple HealthBook can only be considered as a data diary (e.g., for blood pressure) since there are no certified sensors or secure data transmissions are available [77, 81]. Moreover, the plausibility of data provided through these media still has to be validated. "Full-profile search" was also missing in Microsoft HealthVault and Google Health [81]. The study by Jovanov [43] reveals that health data analytics tools have even penetrated the smartwatch industry. In fact, these smartwatches, being the most popular wearable sensors, provide for continuous measurement of physiological parameters including heart rate and temperature. Being in close contact with the skin, these devices can be used to collect biological, environmental, and behavioral information about user activities and potentially provide for user identification.

4.2 Comparative Analysis of Health Data Analytics Tools and Platforms

In this section, a comparative analysis of the reviewed tools and platforms is carried out. The objective is to identify the commonalities and differences between those tools and platforms. This comparative analysis will also allow to uncover the trends in data analytics techniques used in latest tools and platforms. The comparison is based on four main criteria: features and capabilities, data analytics techniques used, types of data used, and the areas of application.

4.2.1 Features and Capabilities

Table 4 shows a comparison between the different tools and platforms described in this section, in terms of features and capabilities.

Based on the feature comparison in Table 4, all the health data analytics tools compared integrate data from various sources, derive new insights from the data, and provide data visualization capabilities. Additionally, most of the tools provide some predictive analysis capabilities that allow healthcare institutions to predict important healthcare parameters and, thus, allow timely decision-making. Some tools are not flexible and do not provide APIs for further development of the data analytics.

4.2.2 Types of Data

In this section, the data types and data sources used by the different tools and platforms are surveyed, and the results are displayed in Table 5. The majority of data analytics tools process semistructured and unstructured data. Moreover, most of the tools integrate data from both internal and external sources.

4.2.3 Areas of Application

Six main areas of application for data analytics in healthcare were identified in Sect. 2. In this section, the application areas of each of the tools and platforms are surveyed and summarized in Table 6. It can be observed that all health data analytics tools are used for healthcare management, that is, analyzing daily operations of the healthcare institution. Moreover, data analytics tools and platforms are increasingly being used for clinical practice in order to improve patient care. It can be seen that IBM Analytics cover the whole spectrum of the application areas and can be used for pharmacovigilance and medical research among others.

4.2.4 Data Analytics Techniques

The data analytics techniques used by each of the tools and platforms are summarized in Table 7. Most of the tools use predictive analytics and machine-learning algorithms in order to derive new insights from the health data. Most tools perform common data analytics tasks such as classification, clustering, and correlation. Moreover, artificial neural networks and decision trees are increasingly being used for data analytics.

Table 4 Feature comparison of health data analytics tools and platforms

Tools and platforms	Data integration (various sources)	Data discovery and analysis	Data visualization/dashboard	Scalable and flexible	Provision of APIs	Predictive analysis/machine learning
Philips HealthSuite Digital Platform	✓	✓	✓	✓	✓	✓
EMIF Platform	✓	✓	✓			✓
Hortonworks	✓	✓	✓			✓
IBM Analytics	✓	✓	✓	✓	✓	✓
HealHub	✓	✓	✓			✓
Sisense	✓	✓	✓	✓	✓	
EHDViz	✓	✓	✓	✓	✓	
Intelligent Care Delivery Analytics platform (ICDA)	✓	✓	✓	✓	✓	✓
EMC Healthcare Analytics	✓	✓	✓	✓		✓
OpenML platform	✓	✓	✓	✓	✓	✓

Table 5 Comparison of data sources and data types

	Data type			Data source	
Tools and platforms	Structured	Semistructured	Unstructured	Internal	External
Philips HealthSuite Digital Platform	✓	✓	✓	✓	✓
EMIF Platform	✓	✓	✓	✓	✓
Hortonworks	✓	✓	✓	✓	✓
IBM Analytics	✓	✓	✓	✓	✓
HealHub	✓			✓	✓
Sisense	✓		✓	✓	✓
EHDViz	✓	✓	✓	✓	✓
Intelligent Care Delivery Analytics platform (ICDA)	✓	✓		✓	
EMC Healthcare Analytics	✓	✓	✓	✓	✓
OpenML platform	✓	✓	✓	✓	✓

Table 6 Comparison of areas of application

Tools and platforms	Disease outbreak and surveillance	Drug–disease association	Pharmaco vigilance	Healthcare management	Clinical/ medical research	Clinical practice
Philips HealthSuite Digital Platform				✓	✓	✓
EMIF Platform				✓		
Hortonworks		✓		✓	✓	
IBM Analytics	✓	✓	✓	✓	✓	✓
HealHub		✓	✓	✓		✓
Sisense				✓		✓
EHDViz				✓		✓
Intelligent Care Delivery Analytics platform (ICDA)				✓		✓
EMC Healthcare Analytics		✓		✓	✓	✓
OpenML platform				✓		✓

4.3 Trends in Health Data Analytics Tools and Platforms

Based on the tools and platforms reviewed in this section, it can be deduced that most health data analytics tools and platforms provide a complete set of services to perform the analytics process, which includes data warehousing, batch processing,

Table 7 Comparison of data analytics techniques

Tools and platforms	Data analytics techniques
Philips HealthSuite Digital Platform	Contextual predictive analytics, decision support algorithms, and machine learning
EMIF Platform	Spearman correlation analysis, hierarchical clustering, K-means clustering, principal component analysis, linear regression, survival analysis, statistical methods
Hortonworks	Combination of Hadoop predictive analytics with a number of data science and iterative machine-learning techniques
IBM Analytics	Predictive analytic solutions comprising of techniques such as artificial neural networks and decision trees
HealHub	Machine-learning techniques
Sisense	Machine-learning techniques, Bots, Natural Language Processing
EMC Healthcare Analytics	Predictive analysis, decision trees, time series, neural networks
OpenML platform	Classification, regression, clustering, data stream classification, learning curve analysis, survival analysis
EHDViz	Predictive analysis, more specifically machine learning and risk algorithms
Intelligent Care Delivery Analytics platform (ICDA)	Predictive analysis (risk assessment analytics)

business intelligence, data workflow orchestration, and machine learning. These tools allow the integration of data from various sources and of different types efficiently without having to learn complex data processing platform such as Hadoop. Indeed, health data analytics tools are integrating data from a wide range of health information systems, including health management systems and electronic health records systems in order to create a unified data model for healthcare institutions. The unified data model is allowing healthcare administrators to have 360° view of the institution.

Moreover, data analytics platforms are increasingly being based on open architectures which allow them to integrate existing infrastructure easily. The open architectures significantly improve the scalability and flexibility of the platforms and support the broadest spectrum of data sources. Much emphasis is also being placed on ensuring high level of security and manageability of data analytics. Health data analytics tools are also focusing on the provision of advanced visualization and dashboards that allow analysts to effortlessly visualize and analyze data from multiple angles without having to depend on developers. The visualization components are providing important features such as interactive analysis, zooming, and drill through capabilities that are allowing greater insight into the health data.

Health data analytics tools and platforms are being used to derive completely new insights that were not possible before. These tools are used to determine adherence to medication, identify high-risk patients, and determine possibility of readmission and morbidity. The use of health data analytics tools is also helping healthcare institutions to predict disease outbreaks, reduce costs of operation, and avoid preventable diseases [71].

5 Use Cases for Health Data Analytics

Healthcare system in many developed countries like the United States is rapidly adopting electronic health records, which is generating a large quantity of electronic health data. Research in health data analytics is also emerging as a new trend in view of improving healthcare service provision. As a result of this advancement, there are unprecedented opportunities to use big data to reduce the costs of healthcare in the United States as well as other developed countries. The following subsections elaborate on some use cases in the area of health data analytics.

5.1 Using Data Analytics to Predict Number of Patients in Hospitals

Hospitals do not have the same number of admissions of patients at all times of the day. One of the major problems consists of deploying the right number of staff at specific times. In order to ensure a very good customer service, ideally hospitals should deploy maximum personnel at peak times and less at other times so as to avoid unnecessary staffing costs. According to a Forbes article (2016), a few hospitals in Paris are using big data analytics to solve the problem of staffing, in view of optimizing costs as well as improving their customer service. These hospitals, which are part of the *Assistance Publique-Hôpitaux de Paris*, have been mining data from a number of internal and external sources – including 10-years' worth of hospital admission records, in order to predict the number of patients at each hospital at given times of the day. They use machine learning to build a predictive model using the past data and predict admission rates at different times of the day. The resulting tool is a web-based application (according to Forbes) that helps healthcare professionals to forecast visits and admission rates for a period of 15 days ahead and plan deployment of staff accordingly. When higher numbers of patients are expected to visit the hospitals, more staff are scheduled in order to reduce waiting times.

5.2 Using Data Analytics to Predict Sepsis Risk and Mortality

Researchers at the University of California Davis have found that routine information such as blood pressure, respiratory rate, temperature, and white blood cell count found in the electronic health records (EHRs) of admitted patients can be used to predict the early stages of sepsis (an immune system response to infection that can damage organs and cause permanent physical and mental disabilities), a leading cause of death and hospitalization in the United States. Moreover, it was also determined that only three parameters, namely lactate level, blood pressure, and respiratory rate, can also predict whether a patient is likely to die from the

disease. They used electronic health records pertaining to 741 adult patients (at the University of California Davis Health System) who met at least two systemic inflammatory response syndrome criteria, to associate patients' vital signs and white blood cell count (WBC) to sepsis occurrence and mortality [31]. Machine-learning algorithms such as generative and discriminative classification (naïve Bayes, support vector machines, Gaussian mixture models, and hidden Markov models) were used to integrate disparate patient data and create a predictive tool for the inference of lactate level and mortality risk.

5.3 Using Data Analytics to Predict Readmissions in Case of Heart Failures

In this study, Bayati et al. [8] have used data from 1172 hospital visits for heart failure to construct a classifier to predict the chance that a patient has of being rehospitalized within 30 days of discharge. The classifier was modeled using data from 793 hospital visits and was tested using data from 379 additional hospital visits. All data were de-identified patient data from the EHR of a hospital. The authors were able to use their model to predict the readmission rate. For this case study, the mean cost of readmissions was 13,679 with a standard error of 1214. The authors claim that if their proposed method would be used, there will be an 18.2% reduction in rehospitalizations and a 3.8% reduction in costs. They also found that the construction and use of predictive models that are custom tailored to populations have a higher accuracy than applications based on the same general rules across different hospitals. The use of local data for the construction and testing of the model also ensured highest predictive performance. Since the readmission patterns and prevalence vary for different hospitals and regions, a model needs to take these into account.

5.4 Using Data Analytics to Predict Decompensation

Decompensation refers to the failure of an organ (especially the liver or heart) to compensate for the functional overload resulting from disease. Often before decompensation, there is a period in which the physiological data of a patient can be used to determine whether she/he is at risk for decompensating [7]. Patients who are critically ill are placed in intensive care units (ICUs) so that they can be closely monitored. With the advancement in technologies, there are also facilities to monitor patients for risk of decompensation in general care units, nursing homes, as well as homes. Some of these technologies were available for many years, such as electrocardiographic monitoring and oxygen monitoring, while others are newer, such as end-tidal monitoring and monitors that allow detection of whether or not a patient is moving. These technologies, however, face the problem of noise which

can raise false alarms. Fortunately, monitors that can compare multiple data streams are also becoming available, and coupled with analytics in the background, it can be determined whether a signal is valid. An example of such a monitor is one that is placed under the mattress of a patient and collects data about the patient's respiratory rate and pulse as well whether the patient is moving or not [12].

6 Challenges in Health Data Analytics

In this section, the challenges for health data analytics are explored. In spite of the several tools and platforms available for health data analytics, there remain a number of challenges that need to be addressed so that the full potential of the available data in healthcare institutions can be harnessed.

6.1 Quality of Health Data

In the context of healthcare, challenges with respect to data quality include inconsistency, irregularity (noise, incompleteness), low standard, unreliability (both primary and secondary data), inaccuracy (or erroneous inclusions), insufficient detail, large volumes, and variety of data and high-dimensionality [50, 69]. As identified by Powell et al. [69], the success depends on "high-quality" data provided by different sources. Over the past two decades, data quality has attracted much research including Total Data Quality Management (TDQM). Several challenges related to quality of data and information include both technical challenges, such as data integration from multiple sources, and nontechnical challenges, such as ensuring the right recipient receives the right data/information in the right format at the right place and at the right time [57]. The quality of data has a direct impact on the effectiveness of the operations being performed on data, such as capture, storage, searching, sharing, analysis, and visualization [51]. Low-quality data, such as missing data, inaccurate, or incomplete data, disturb the proper functioning of the different processes. Additionally, irregularity of data incorporates both noisiness and incompleteness. Noise is often incorporated in data itself where data cannot be interpreted and are meaningless. Healthcare data are often incomplete as not all data are being recorded properly [50].

The high volume of data generated from internal and external sources gives rise to a large number of data elements stored in them, leading to high dimensionality when dividing the healthcare data [50]. Processes like data mining, therefore, become difficult in healthcare applications due to corrupted, inconsistent, missing, or unstandardized data. For example, information might be recorded in different formats in varying data sources [10, 22, 47]. Powell et al. [69] provided two scenarios showing data quality issues in healthcare. In the first scenario, there was clinical evidence that diagnosis for only a few diseases was detected due to

incomplete data. In the second scenario, accurate data were not provided in all cases; for example, there were accurate data for tumor registries but inaccurate data for outpatient treatment [69].

In the work by Berman et al. [10] related to the Protein Data Bank, new data are integrated in the existing data archive. It is, therefore, important that standard formats are used. This work relates to issues related to standardization and integration. Two approaches have been used in this work to address this issue: file-by-file analysis and batch processing [10]. Edwards et al. [22] further extend on the issue related to standardization in terms of representation of data. To allow efficient data storage, it is imperative that healthcare data be in a standardized format and make use of standard vocabularies for purposes of concept representation and communication [22].

6.2 Infrastructural Challenges

A fault-tolerant and scalable physical infrastructure is vital for the operation of a big data analytics project. This infrastructure is usually based on a distributed model, where the data can be physically stored in various locations connected through high-speed networks [54]. Demchenko et al. [20] identified the following infrastructure requirements for emerging big data analytics projects:

- Support the running of long experiments with large data volumes produced at high speed
- Confidentiality, integrity, and accountability of data
- Multitier interlinked data distribution and replication
- On-demand infrastructure provisioning
- Trusted environment for data storage and processing

AbuKhousa et al. [1] identified several challenges faced by an e-health cloud infrastructure which inherits the challenges of processing sensitive medical data of health information systems and cloud computing. As most health data analytics platforms use cloud computing technologies, the underlying infrastructures for these platforms will, therefore, face challenges such as availability, data management, scalability, flexibility, interoperability, security, and privacy.

6.3 Data Analytic Algorithms and Approaches

As discussed in Sect. 2, data analytics can be applied to a variety of different areas in healthcare, namely drug–drug association, disease outbreak detection and surveillance, pharmacovigilance, healthcare management, clinical/medical research, and clinical practice. However, these areas are very diverse with some directly and others indirectly related to the medical field. Hence, different approaches are used by each.

Newer methods, such as data mining, Bayesian statistics, optimization modeling, social network analysis, and agent-based simulation, have been combined to the previously well-established techniques, such as biostatistics and epidemiologic analysis, causal modeling, and Monte Carlo and discrete-event simulation [86].

In their review paper on data mining of big data in health analytics, Herland et al. [38] have categorized health informatics into different levels, and for each of these, they have provided a set of tools that can be used for the analysis [4, 13, 25, 32, 65, 75, 89]. Belle et al. [9] have also classified big data in medicine under different areas and have provided a set of tools for each. They have proposed the MapReduce framework, support vector machines (SVM), and wavelet analysis. As it can be observed, several methods have been proposed since the data analytics methods are very problem specific, and there is a need to combine data from disparate sources, for example, combining patient data, physician profile, and environmental variable features, to improve the management of risk-stratified patients to receive the most appropriate care [55].

The criteria for the evaluation of a big data analytics platform for healthcare include availability, continuity, ease of use, scalability, ability to manipulate at different levels of granularity, privacy and security enablement, and quality assurance [71]. Since healthcare data are rarely standardized, one major challenge of data analytics algorithms in healthcare is the need to deal with incompatible formats of highly fragmented data. Moreover, unstructured textual data like clinical notes can be difficult to understand in the right context [70]. Real-time big data analytics are key requirements for healthcare. Much research efforts are geared toward continuous data acquisition and cleansing and addressing the lag between data collection and processing [71].

Some data mining methods may yield good performance for one type of problems, while other methods are more appropriate for other types of problems [58]. To decide on an appropriate algorithm, it is important to understand the appropriateness of the method and its performance for a specific problem. The chosen algorithm needs to be extensively evaluated and experimented before applying it in a real medical system. There is no such algorithm which is best suited for all problems in the medical domain, since each algorithm is applicable to specific problems [58].

6.4 Security and Privacy Challenges

Health data analytics have enormous potential to reduce cost of operations, improve clinical practice, and improve the quality of the overall healthcare. However, this opportunity raises a series of security, privacy, ethical, and legal challenges [17]. Hence, we have the concept of Protected Health Information (PHI) which refers to individually identifiable health information transmitted by electronic media, maintained in electronic media, or transmitted or maintained in any other form or medium [63].

Security and privacy concerns related to health data that are accessed through remote locations such as cloud present a significant barrier to the adoption of big data analytics in the health sector. Some institutions have been experimenting with private clouds, which are, however, limited in terms of scalability. The concept of Cloud Service Providers has emerged as potential hosting spaces, and they provide additional security measures on top of the cloud service. In the long run, they can be better than private clouds as they can invest in measures against hacking and cyberattacks.

Often, identifiable health information cannot be disclosed without patient consent. In many cases, it is not practical to obtain individual patient consent because of the very large data sets involved. One approach to facilitate the use of health information for the purposes of predictive analytics is to de-identify data. The Privacy Rule of Health Insurance Portability and Accountability Act (HIPAA) of 1996 establishes minimum Federal standards for protecting the privacy of individually identifiable health information. The Privacy Rule establishes conditions under which covered entities can provide access to and use of PHI. The Privacy Rule permits covered entities to use and disclose data that have been de-identified without obtaining an authorization and without further restrictions on use or disclosure because de-identified data are no longer PHI and, therefore, are not subject to the Privacy Rule (https://www.hhs.gov/hipaa/for-professionals/privacy/index.html).

Predictive analytics in medicine can effectively help in identifying patients at high or low risk for serious complications, thus optimally allocating scarce clinical resources. Predictive analytics models make treatment recommendations that are designed to improve overall health outcomes among all patients, and these recommendations may conflict with physicians' ethical obligations to act in the best interests of individual patients. Health practitioners should be able to override computerized recommendations when they have sound reasons to believe that the predictive model did not capture some considerations. Such exceptions would allow treating physicians to play their traditional role as patient advocates.

The potential of health data analytics cannot be realized without collecting and analyzing vast amounts of heterogeneous data [49]. Moreover, the data involved may not all be owned or controlled by a single organization or institution. Sharing data might enable various organizations to identify and fill the gaps across their common coverage area, such as patterns of inappropriate use of antibiotic medications. However, competing healthcare institutions that treat common patients may be reluctant to share relevant data, fearing that others could use its data for their advantage.

7 Future Directions

Despite the enormous progress made in the field of health data analytics, there are still several challenges that need to be addressed so that healthcare institutions can leverage the full benefits of their accumulated data. Based on the studies reviewed in

this chapter, it can be clearly observed that health data analytics has not yet matured and there is still a need for focused research. Several research avenues for health data analytics are, therefore, defined in this section in an attempt to provide researchers and practitioners with future directions in this domain.

7.1 Health Internet of Things

The Internet of Things is having a tremendous impact in the healthcare sector with 40% of IoT-related technology forecasted to be health related by 2020 [21]. Wearable IoT integrates wearable sensors that monitor human factors such as health, wellness, behaviors, and other useful information that have the potential to allow individuals to better manage their health [39]. However, a number of technical challenges need to be addressed in order for the Wearable Internet of Things (WIoT) to achieve multidimensional success. These include the generation of a flexible and robust framework for networking, storage, computation, and visualization while designing healthcare solutions which are clinically accepted and operational.

Hassanalieragh et al. [34] identified some challenges with respect to analytics with wearable sensor data. First, they argued that analytics on data from the sensors is challenging since there is a need to cope with streaming data with a number of missing values and varying dimensions and semantics of data. Although the machine-learning algorithms have matured, they are, however, not designed to deal with time-varying feature dimensionality, and incomplete data vectors, which if not catered for properly, can affect classification performance. Second, while there is an enormous amount of sensor data, these are completely untagged and need to be matched with the physician diagnoses. However, with the high load of physicians, this activity is nearly unfeasible. Third, sensor data and historical information available in clinical records are very different in nature. This heterogeneity is a major challenge for conventional machine-learning approaches that work primarily with homogeneous data.

According to Riazul Islam et al. [73], there is a need for a customized computing platform for the IoT-based healthcare systems. Moreover, they argue that libraries, for example, a specific class of disease-oriented libraries, and appropriate frameworks should be customized so that software developers for healthcare systems can make effective use of given classes, documents, templates, codes, and other useful data. A number of other limitations exist in the available platforms supporting data analytics for healthcare [71], and these should be addressed in order to ensure a large-scale adoption. Data analytics in healthcare platforms should be menu-driven, transparent, and user-friendly. Algorithms, models, and methods should be dynamically and easily accessible through drop-down menus. Since real-time analytics are vital in healthcare, the time lag between data collection and processing should be addressed and continuous data acquisition and cleansing should be favored. Other issues like ownership of data, incompatible and fragmented data, and standardization of data should also be addressed.

7.2 Precision Medicine

Precision medicine is an emerging approach for individualizing the practice of medicine [61] based on information about health and disease at the molecular, cellular, and organ levels [26]. It is also known as personalized, predictive, preventive, and participatory (P4) medicine [40]. Two major computational challenges [28] for precision medicine include the developing algorithms for, first, individual phenotyping, which refers to the annotation of patient records with disease conditions, and, second, the integration of electronic health records data with omics data in order to better understand the disease and its treatments. These challenges are mainly due to the noisiness, incompleteness, heterogeneity, and different formats of the electronic health records and genomic data.

There are a number of significant computational advances in precision medicine such as cloud-based toolkits and workflow platforms that provide for high-throughput processing and analysis of omics data [2]. Additionally, graphics processing units (GPUs), which provide faster computations compared to central processing units (CPUs), are being exploited to handle the exponentially growing data. However, given the heterogeneous nature of omics data and challenges in communications and synchronizations, future works are required to develop parallelization algorithms. Moreover, there is a need to create lightweight programming environments with a number of cloud-based utilities and to validate the reliability of the platforms before a large-scale adoption can be envisaged.

7.3 Data Analytics for Evidence-Based Medicine and Drug Repurposing

Evidence-based medicine (EBM) has, for a long time, been very successful in clinical decision-making in many countries [30]. It is the process of integrating individual clinical expertise, patients' preferences, and evidence from randomized clinical trials and research findings. The ethical care of the patient is a top priority for the success of EBM, and patients now demand better individualized evidence, presented and explained in a personalized way. However, in recent years, the implementation of evidence-based practice has become a major challenge, one of the reasons being that the volume of evidence is growing at an unexpected rate. Therefore, much research effort should be geared toward health data analytics techniques to extract the best relevant evidence from the ever-increasing volumes of clinical guidelines and research studies, for patients on a case-to-case basis, for better clinical decision-making. A major challenge for these techniques is to deal with data sets which are not only large but also usually complex, heterogeneous, high-dimensional, time-varying, noisy, and weakly structured or unstructured.

Drug repurposing, also known as drug repositioning, is the use of existing drugs to treat new diseases. In the past decades, this has become an increasingly

important strategy for new drug discovery, given the very high cost and failure rate of new drug development [88]. With the increasing growth of drug-related data, new computational strategies and techniques for drug repositioning are emerging [52]. These methods integrate data from various sources. However, Li et al. [52] identify a few issues which need to be addressed. Factors like missing data, data bias, and technical limitations of computational methods need to be considered when applying the computational models into practical use. Moreover, the performance evaluation of the proposed models is quite hard, because of the lack of structured standard for drug repositioning. Gligorijević et al. [28] suggest the utilization of Topological Data Analysis methods to deal with the high dimensionality (volume) of biomedical data, the so-called anytime-algorithms to deal with the velocity of biomedical data, and Matrix Factorization-based methods to deal with the heterogeneity (variety) of biomedical data.

7.4 Improved Predictive Analytics

Predictive analytics have been used in different areas of healthcare as discussed in Sect. 2. According to Raghupathi and Raghupathi [71], big data analytics and applications in the field of healthcare are still at a nascent stage of development. However, they claim that this area is developing rapidly with the advances of new platforms and tools. There are still some healthcare areas where the full potential of predictive and big data analytics have not been exploited. Some of the areas are described further.

7.4.1 Identification and Management of High-Cost and High-Risk Patients

According to Bates et al. [7], the outcomes of predictive modeling and analytics currently come mostly from low-risk patients. It is, therefore, of utmost importance to apply concepts of predictive analytics for high-risk and high-cost patients' identification and management. Parikh et al. [67] also support this point by stating that existing systems have not progressed much in the field of risk stratification. As a future direction, the authors suggest that predictions models be integrated with clinical systems to assist health professionals to make decisions.

7.4.2 Readmissions of Patients

Readmissions of patients have been costly to healthcare institutions where patients get admitted again after receiving care at the particular institutions. Some work has already been done in the field [3, 8, 11, 44, 78, 92]. Bates et al. [7] identify

readmissions of patients as a use case where additional research is required using analytics and big data to perform readmission risk prediction.

7.4.3 Triage of Patients

Another use case identified by Bates et al. [7] where there is potential for further research is triage, where risk complications are identified when a patient first visits a hospital. Few models exist in this domain. Sun et al. [79] present a predictive model where the need for a patient to be admitted in the emergency department (ED) is determined at the time of triage. However, the use of big data analytics has not been exploited by the model.

7.4.4 Treatment Optimization for Diseases Affecting Multiple Organ Systems

Bates et al. [7] argue that current approaches of Big Data Analytics focus more on use of analytics for patients with one condition rather than multiple conditions. Shams et al. [78] have proposed a hybrid prediction model that integrates classification and timing-based analytics models for patients suffering from four different conditions, though it has certain limitations. They plan to improve on their model in future.

7.5 Real-Time Health Data Analytics

While a number of real-time and stream analytics applications exist in healthcare [45, 48, 80, 85, 91], not all of them have been tested in real clinical conditions or with real data sets [45, 85]. Zhang et al. [91] implemented a Health Data Stream Analytics System which was still under evaluation at the time of publication. Future directions included proposed "formal clinical trials to evaluate the performance." Further research in this area is still required. In the context of real-time monitoring of patients and resources, it is important to integrate predictions with clinical systems "to help physicians and other healthcare professionals make decisions and track real-time quality" [67]. Future research directions in real-time healthcare analytics should not be disjoint from the challenges related to large volume streams, continuous data collection, cleansing, processing, near-real-time responsiveness, accuracy, and the impact on performance and scalability [1, 71, 85]. Moreover, although the technical aspects regarding the provision of real-time healthcare analytics might be satisfied, poor quality of data can still be a barrier. For instance, in some cases, this can be hindered by the unreliability, incompleteness, or unavailability of the relevant data. Therefore, data quality can be added to the research agenda related to real-time healthcare analytics.

7.6 New Data Analytics for Clinical and Medical Research

In the field of clinical and medical research, new data analytics techniques are being used extensively for the better understanding of diseases like cancer, heart diseases, and others. canSAR (http://cansar.icr.ac.uk) is a cancer-focused knowledge base, which has been developed to support cancer research and translate findings from fundamental research into medical practice and meaningful health outcomes as well as in drug discovery. canSAR integrates data from diverse sources like data on the genome, data on proteins, pharmacological data, drug and chemical data, as well as structural biology, and protein networks data. As on December 11, 2018, canSAR contains 556,825 proteins from 2148 organisms, data for 12,172 cancer cell and nontransformed cell line models, 6,367,677 experimental data points from patient-derived tissue samples, and 145,978 3D structures (http://cansar.icr.ac.uk/cansar/data-sources/) and is still growing. Although the system is already providing a number of functionalities, in the next phase it aims at improving the search and browsing power and developing expert tools [83]. The knowledge base aims at becoming, in the future, a system where scientists can simply ask a question pertaining to cancer and receive an answer very quickly.

Researchers are also using big data analytics on cancer data to investigate why some cancer patients relapse and others do not, and some initial results have already been published by Pan et al. [66], but there is still a lot that can be done. Yet, another work in the field of cancer and big data is the setting up of a database of cancer genome mutations to quickly identify which mutations in a tumor are most important. The project involves scanning cancer literature and performing constant database updates. The next step will involve crowdsourcing the data from more scientists who are willing to participate. Khurana et al. [46] have created a computational tool, FunSeq, that can mine terabytes of disregarded genomic data to find possible unknown drivers of cancer.

8 Conclusion

With the explosion of data being provided at unprecedented speed from multiple sources, the need for predictive analytics is being explored in various key sectors including healthcare. Predictive analytics can be used in conjunction with big data to allow valuable insights to be derived to enhance clinical practice and patients' outcomes as well as lower healthcare costs in conjunction with personalized medicine. This chapter visits some application areas in the healthcare sector which have been classified into drug–disease association, disease outbreak detection and surveillance, pharmacovigilance, healthcare management, clinical research, and clinical practice. Data analytics can be applied to both internal and external sources of data to gain insights on diseases trends, effectiveness of treatments, cost-effectiveness of operations, proper management of resources, performance of

individual departments, disease outbreaks around the world, behavioral patterns of patients, and health events in different countries among others. The main sources of healthcare data, both internal and external, have been identified and analyzed.

Furthermore, different tools and platforms available in the healthcare sector have been analyzed. A comparative analysis is carried out for the different tools and platforms in terms of features and capabilities, types of data incorporated, data sources being integrated, areas of application, and data analytics techniques used. The main findings of the comparative study are that most health data analytics tools and platforms provide a complete set of services to perform the analytics process. Additionally, these tools integrate data from a wide range of health information systems, including health management systems and electronic health record systems, in order to create a unified data model for healthcare institutions. Data analytics platforms are increasingly being based on open architectures, which allow integration into existing infrastructure. These platforms emphasize on the provision of advanced visualization and dashboards that allow analysts to effortlessly visualize and analyze data from multiple angles without relying on technical experts. Health data analytics tools and platforms are being used to derive completely new insights that were not possible before.

Finally, the chapter presents a number of challenges that need to be addressed in terms of quality of health data, infrastructural challenges, data analytics algorithms and approaches, and security and privacy challenges so that the full potential of predictive analytics can be applied to healthcare. Future directions for health data analytics are discussed with respect to IoT healthcare and precision medicine, data analytics for evidence-based medicine and drug repurposing, improved predictive analytics, real-time health data analytics, and new data analytics techniques for clinical and medical research. More research needs to be carried out in these areas so that healthcare institutions and patients can leverage maximum benefits. The main contribution of this chapter is an attempt to review existing work in the area of health data analytics. It is expected that this review will help researchers and practitioners to shape new research directions to address the challenges in this domain.

Funding This manuscript is part of the work of a funded project by the Mauritius Research Council, Reference HPCRIG-A03.

References

1. E. AbuKhousa, N. Mohamed, J. Al-Jaroodi, e-Health cloud: Opportunities and challenges. Future Internet **4**(4), 621–645 (2012)
2. A. Alyass, M. Turcotte, D. Meyre, From big data analysis to personalized medicine for all: Challenges and opportunities. BMC Med. Genomics **8**(1), 33 (2015)
3. R. Amarasingham, R.E. Patzer, M. Huesch, N.Q. Nguyen, B. Xie, Implementing electronic health care predictive analytics: Considerations and challenges. Health Aff. (Project Hope) **33**(7), 1148–1154 (2014)
4. J. Annese, The importance of combining MRI and large-scale digital histology in neuroimaging studies of brain connectivity and disease. Front. Neuroinform. **6**, 13 (2012)

5. M.A. Badgeley, K. Shameer, B.S. Glicksberg, M.S. Tomlinson, M.A. Levin, P.J. McCormick, A. Kasarskis, D.L. Reich, J.T. Dudley, EHDViz: Clinical dashboard development using open-source technologies. BMJ Open **6**(3), e010579 (2016)
6. I. Bardhan, J. Oh, Z. Zheng, K. Kirksey, Predictive analytics for readmission of patients with congestive heart failure. Inf. Syst. Res. **26**(1), 19–39 (2015). https://doi.org/10.1287/isre.2014.0553
7. D.W. Bates, S. Saria, L. Ohno-Machado, A. Shah, G. Escobar, Big data in health care: Using analytics to identify and manage high-risk and high-cost patients. Health Aff. (Project Hope) **33**(7), 1123–1131 (2014)
8. M. Bayati, M. Braverman, M. Gillam, K.M. Mack, G. Ruiz, M.S. Smith, E. Horvitz, Data-driven decisions for reducing readmissions for heart failure: General methodology and case study. PLoS One **9**(10), e109264 (2014)
9. A. Belle, R. Thiagarajan, S.M.R. Soroushmehr, F. Navidi, D.A. Beard, K. Najarian, Big data analytics in healthcare. Biomed. Res. Int. **2015**, 370194 (2015)
10. H.M. Berman, T.N. Bhat, P.E. Bourne, Z. Feng, G. Gilliland, H. Weissig, J. Westbrook, The Protein Data Bank and the challenge of structural genomics. Nat. Struct. Biol. **7**(Suppl), 957–959 (2000)
11. J. Billings, J. Dixon, T. Mijanovich, D. Wennberg, Case finding for patients at risk of readmission to hospital: Development of algorithm to identify high risk patients. BMJ **333**(7563), 327 (2006)
12. H. Brown, J. Terrence, P. Vasquez, D.W. Bates, E. Zimlichman, Continuous monitoring in an inpatient medical-surgical unit: A controlled clinical trial. Am. J. Med. **127**(3), 226–232 (2014)
13. A.J. Campbell, J.A. Cook, G. Adey, B.H. Cuthbertson, Predicting death and readmission after intensive care discharge. Br. J. Anaesth. **100**(5), 656–662 (2008)
14. K.C.C. Chan, Big data analytics for drug discovery, in *2013 IEEE International Conference on Bioinformatics and Biomedicine*, (IEEE, Piscataway, 2013), pp. 1–1
15. E.S. Chen, G. Hripcsak, H. Xu, M. Markatou, C. Friedman, Automated acquisition of disease drug knowledge from biomedical and clinical documents: An initial study. J. Am. Med. Inform. Assoc. **15**(1), 87–98 (2008)
16. A.P. Chiang, A.J. Butte, Systematic evaluation of drug-disease relationships to identify leads for novel drug uses. Clin. Pharmacol. Ther. **86**(5), 507–510 (2009)
17. I.G. Cohen, R. Amarasingham, A. Shah, B. Xie, B. Lo, The legal and ethical concerns that arise from using complex predictive analytics in health care. Health Aff. (Project Hope) **33**(7), 1139–1147 (2014)
18. A. Culotta, Towards detecting influenza epidemics by analyzing Twitter messages, in *Proceedings of the First Workshop on Social Media Analytics – SOMA '10*, (ACM Press, New York, 2010), pp. 115–122
19. Dell EMC Data Analytic Services for Healthcare [Online]. Available from: https://www.ciosummits.com/Online_Assets_Dell_EMC_Service_Overview_-_Chicago.pdf. Accessed 5 Sept 2019
20. Y. Demchenko, P. Grosso, C. de Laat, P. Membrey, Addressing big data issues in Scientific Data Infrastructure, in *2013 International Conference on Collaboration Technologies and Systems (CTS)*, (IEEE, San Diego, California, 2013), pp. 48–55
21. D.V. Dimitrov, Medical internet of things and big data in healthcare. Healthc. Inform. Res. **22**(3), 156–163 (2016)
22. J.R. Edwards, D.A. Pollock, B.A. Kupronis, W. Li, J.S. Tolson, K.D. Peterson, R.B. Mincey, T.C. Horan, Making use of electronic data: The National Healthcare Safety Network eSurveillance Initiative. Am. J. Infect. Control **36**(3 Suppl), S21–S26 (2008)
23. EMC Healthcare Analytics Solution [Online]. Available at: https://www.emc.com/collateral/solution-overview/h11357-healthcare-analytics-so.pdf. Accessed 5 Sept 2019
24. EMIF, European Medical Information Framework [Online]. Available at: http://www.emif.eu/. Accessed 5 Sept 2019
25. F. Estella, B.L. Delgado-Marquez, P. Rojas, O. Valenzuela, B.S. Roman, I. Rojas, Advanced system for automously classify brain MRI in neurodegenerative disease, in *2012 International Conference on Multimedia Computing and Systems*, (IEEE, Piscataway, 2012), pp. 250–255

26. M. Flores, G. Glusman, K. Brogaard, N.D. Price, L. Hood, P4 medicine: How systems medicine will transform the healthcare sector and society. Pers. Med. **10**(6), 565–576 (2013)
27. R. Frijters, M. van Vugt, R. Smeets, R. van Schaik, J. de Vlieg, W. Alkema, Literature mining for the discovery of hidden connections between drugs, genes and diseases. PLoS Comput. Biol. **6**(9), e1000943 (2010)
28. V. Gligorijević, N. Malod-Dognin, N. Pržulj, Integrative methods for analyzing big data in precision medicine. Proteomics **16**(5), 741–758 (2016)
29. D. Gotz, H. Stavropoulos, J. Sun, F. Wang, ICDA: A platform for intelligent care delivery analytics. AMIA Annu. Symp. Proc. **2012**, 264–273 (2012)
30. T. Greenhalgh, J. Howick, N. Maskrey, Evidence Based Medicine Renaissance Group, Evidence based medicine: A movement in crisis? BMJ **348**, g3725 (2014)
31. E. Gultepe et al., From vital signs to clinical outcomes for patients with sepsis: A machine learning basis for a clinical decision support system. J. Am. Med. Inform. Assoc. **21**(2), 315–325 (2013)
32. T. Haferlach, A. Kohlmann, L. Wieczorek, G. Basso, G.T. Kronnie, M.-C. Béné, J. De Vos, J.M. Hernández, W.-K. Hofmann, K.I. Mills, A. Gilkes, S. Chiaretti, S.A. Shurtleff, T.J. Kipps, L.Z. Rassenti, A.E. Yeoh, P.R. Papenhausen, W.-M. Liu, P.M. Williams, R. Foà, Clinical utility of microarray-based gene expression profiling in the diagnosis and subclassification of leukemia: Report from the International Microarray Innovations in Leukemia Study Group. J. Clin. Oncol. **28**(15), 2529–2537 (2010)
33. R. Harpaz, W. DuMouchel, M. Schuemie, O. Bodenreider, C. Friedman, E. Horvitz, A. Ripple, A. Sorbello, R.W. White, R. Winnenburg, N.H. Shah, Toward multimodal signal detection of adverse drug reactions. J. Biomed. Inform. **76**, 41–49 (2017)
34. M. Hassanalieragh, A. Page, T. Soyata, G. Sharma, M. Aktas, G. Mateos, B. Kantarci, S. Andreescu, Health monitoring and management using Internet-of-Things (IoT) sensing with cloud-based processing: Opportunities and challenges, in *2015 IEEE International Conference on Services Computing*, (IEEE, New York, 2015), pp. 285–292
35. Healthcare Analytics & Business Intelligence | Sisense [Online]. Available at: https://www.sisense.com/solutions/healthcare/. Accessed 10 Sept 2019
36. Healthcare Predictive Analytics – Big Data Analytics in Medicine and Genomics | Hortonworks [Online]. Available at: https://hortonworks.com/solutions/healthcare/. Accessed 10 Sept 2019
37. HealthSuite Ecosystem | Philips Healthcare [Online]. Available at: https://www.usa.philips.com/healthcare/innovation/about-health-suite. Accessed 10 Sept 2019
38. M. Herland, T.M. Khoshgoftaar, R. Wald, A review of data mining using big data in health informatics. J. Big Data **1**(1), 2 (2014)
39. S. Hiremath, G. Yang, K. Mankodiya, Wearable Internet of Things: Concept, architectural components and promises for person-centered healthcare, in *2014 4th International Conference on Wireless Mobile Communication and Healthcare – Transforming Healthcare Through Innovations in Mobile and Wireless Technologies (MOBIHEALTH)*, (IEEE, Piscataway, 2014), pp. 304–307
40. L. Hood, S.H. Friend, Predictive, personalized, preventive, participatory (P4) cancer medicine. Nat. Rev. Clin. Oncol. **8**(3), 184–187 (2011)
42. P.L. Ingrassia, L. Carenzo, F.L. Barra, D. Colombo, L. Ragazzoni, M. Tengattini, F. Prato, A. Geddo, F. Della Corte, Data collection in a live mass casualty incident simulation: Automated RFID technology versus manually recorded system. Eur J Emerg Med **19**(1), 35–39 (2012)
43. E. Jovanov, Preliminary analysis of the use of smartwatches for longitudinal health monitoring, in *Conference Proceedings: Annual International Conference of the IEEE Engineering in Medicine and Biology Society 2015*, (IEEE, Piscataway, 2015), pp. 865–868
44. D. Kansagara, H. Englander, A. Salanitro, D. Kagen, C. Theobald, M. Freeman, S. Kripalani, Risk prediction models for hospital readmission: A systematic review. J. Am. Med. Assoc. **306**(15), 1688–1698 (2011)

45. H. Khazaei, N. Mench-Bressan, C. McGregor, J.E. Pugh, Health informatics for neonatal intensive care units: An analytical modeling perspective. IEEE J. Transl. Eng. Health Med. **3**, 3000109 (2015)
46. E. Khurana, Y. Fu, V. Colonna, X.J. Mu, H.M. Kang, T. Lappalainen, A. Sboner, L. Lochovsky, J. Chen, A. Harmanci, J. Das, A. Abyzov, S. Balasubramanian, K. Beal, D. Chakravarty, D. Challis, Y. Chen, D. Clarke, L. Clarke, F. Cunningham, M. Gerstein, Integrative annotation of variants from 1092 humans: Application to cancer genomics. Science **342**(6154), 1235587 (2013)
47. H.C. Koh, G. Tan, Data mining applications in healthcare. J. Healthc. Inf. Manage. **19**(2), 64–72 (2005)
48. E.I. Konstantinidis, A.S. Billis, L. Plotegher, G. Conti, P.D. Bamidis, Indoor location IoT analytics "in the wild": Active and healthy ageing cases, in *XIV Mediterranean Conference on Medical and Biological Engineering and Computing 2016. IFMBE Proceedings*, ed. by E. Kyriacou, S. Christofides, C. S. Pattichis, (Springer International Publishing, Cham, 2016), pp. 1231–1236
49. C.S. Kruse, R. Goswamy, Y. Raval, S. Marawi, Challenges and opportunities of big data in health care: A systematic review. JMIR Med. Inform. **4**(4), e38 (2016)
50. V. Kumar, H. Park, R. Basole, M. Braunstein, M. Kahng, D. Chau, A. Tamersoy, D. Hirsh, N. Serban, J. Bost, B. Lesnick, B. Schissel, M. Thompson, Exploring clinical care processes using visual and data analytics: Challenges and opportunities, in 20th ACM SIGKDD Conference on Knowledge Discovery and Data Mining Workshop on Data Science for Social Good, 2014
51. LexisNexis, Why data quality is the greatest challenge and opportunity for health care (2016). Available at: https://www.lexisnexis.com/risk/downloads/whitepaper/Data-Quality-POV.pdf
52. J. Li, S. Zheng, B. Chen, A.J. Butte, S.J. Swamidass, Z. Lu, A survey of current trends in computational drug repositioning. Brief. Bioinform. **17**(1), 2–12 (2016)
53. D. Lopez, M. Gunasekaran, B.S. Murugan, H. Kaur, K.M. Abbas, Spatial big data analytics of influenza epidemic in Vellore, India, in *IEEE International Conference on Big Data 2014*, (IEEE, Washington, DC, USA, 2014), pp. 19–24
54. D. Luna, J.C. Mayan, M.J. García, A.A. Almerares, M. Househ, Challenges and potential solutions for big data implementations in developing countries. Yearb. Med. Inform. **9**, 36–41 (2014)
55. G. Luo, B.L. Stone, F. Sakaguchi, X. Sheng, M.A. Murtaugh, Using computational approaches to improve risk-stratified patient management: Rationale and methods. JMIR Res. Protoc. **4**(4), e128 (2015)
56. Z. Lv, J. Chirivella, P. Gagliardo, Bigdata oriented multimedia mobile health applications. J. Med. Syst. **40**(5), 120 (2016)
57. S.E. Madnick, R.Y. Wang, Y.W. Lee, H. Zhu, Overview and framework for data and information quality research. J. Data Inf. Qual. **1**(1), 1–22 (2009)
58. P. Mahindrakar, D. Hanumanthappa, Data mining in healthcare: A survey of techniques and algorithms with its limitations and challenges. Int. J. Eng. Res. Appl. **3**(6), 937–941 (2013)
60. MEDLINE Fact Sheet [Online]. Available at: https://www.nlm.nih.gov/pubs/factsheets/medline.html. Accessed 10 Sept 2019
61. R. Mirnezami, J. Nicholson, A. Darzi, Preparing for precision medicine. N. Engl. J. Med. **366**(6), 489–491 (2012)
62. B. Moore, The potential use of radio frequency identification devices for active monitoring of blood glucose levels. J. Diabetes Sci. Technol. **3**(1), 180–183 (2009)
63. NIH HIPAA Privacy Rule and Its Impacts on Research [Online]. Available at: https://privacyruleandresearch.nih.gov/pr_07.asp. Accessed 10 Sept 2019

64. A. Nikfarjam, A. Sarker, K. O'Connor, R. Ginn, G. Gonzalez, Pharmacovigilance from social media: Mining adverse drug reaction mentions using sequence labeling with word embedding cluster features. J. Am. Med. Inform. Assoc. **22**(3), 671–681 (2015)
65. I. Ouanes, C. Schwebel, A. Français, C. Bruel, F. Philippart, A. Vesin, L. Soufir, C. Adrie, M. Garrouste-Orgeas, J.-F. Timsit, B. Misset, Outcomerea Study Group, A model to predict short-term death or readmission after intensive care unit discharge. J. Crit. Care **27**(4), 422.e1–422.e9 (2012)
66. H. Pan, Y. Jiang, M. Boi, F. Tabbò, D. Redmond, K. Nie, M. Ladetto, A. Chiappella, L. Cerchietti, R. Shaknovich, A.M. Melnick, G.G. Inghirami, W. Tam, O. Elemento, Epigenomic evolution in diffuse large B-cell lymphomas. Nat. Commun. **6**, 6921 (2015)
67. R.B. Parikh, M. Kakad, D.W. Bates, Integrating predictive analytics into high-value care: The dawn of precision delivery. J. Am. Med. Assoc. **315**(7), 651–652 (2016)
68. G. Perna, Shannon Medical Center. Leaders at one Texas-based hospital are not taking hand hygiene for granted. Healthc. Inform. **30**(1), 28–30 (2013)
69. A.E. Powell, H.T.O. Davies, R.G. Thomson, Using routine comparative data to assess the quality of health care: Understanding and avoiding common pitfalls. Qual. Saf. Health Care **12**(2), 122–128 (2003)
70. K. Priyanka, N. Kulennavar, A survey on big data analytics in health care. Int. J. Comput. Sci. Inf. Technol. **5**(4), 5865–5868 (2014)
71. W. Raghupathi, V. Raghupathi, Big data analytics in healthcare: Promise and potential. Health Inf. Sci. Syst. **2**, 3 (2014)
72. D.C. Ranasinghe, R.L. Shinmoto Torres, A.P. Sample, J.R. Smith, K. Hill, R. Visvanathan, Towards falls prevention: A wearable wireless and battery-less sensing and automatic identification tag for real time monitoring of human movements, in *Conference Proceedings: Annual International Conference of the IEEE Engineering in Medicine and Biology Society 2012*, (IEEE, Piscataway, 2012), pp. 6402–6405
73. S.M. Riazul Islam, D. Kwak, M. Humaun Kabir, M. Hossain, K.-S. Kwak, The internet of things for health care: A comprehensive survey. IEEE Access **3**, 678–708 (2015)
74. Sagitec HealHub [Online]. Available at: http://www.sagitec.com/healhub. Accessed 10 Sept 2019
75. R. Salazar, P. Roepman, G. Capella, V. Moreno, I. Simon, C. Dreezen, A. Lopez-Doriga, C. Santos, C. Marijnen, J. Westerga, S. Bruin, D. Kerr, P. Kuppen, C. van de Velde, H. Morreau, L. Van Velthuysen, A.M. Glas, L.J. Van't Veer, R. Tollenaar, Gene expression signature to improve prognosis prediction of stage II and III colorectal cancer. J. Clin. Oncol. **29**(1), 17–24 (2011)
76. A. Sarker, R. Ginn, A. Nikfarjam, K. O'Connor, K. Smith, S. Jayaraman, T. Upadhaya, G. Gonzalez, Utilizing social media data for pharmacovigilance: A review. J. Biomed. Inform. **54**, 202–212 (2015)
77. E.G. Schulz, C.L. Neumann, Interventional decentralized telemonitoring: Bridging the gap between patient's device and physician's needs in well selected indications. Kidney Blood Press. Res. **40**(2), 130–140 (2015)
78. I. Shams, S. Ajorlou, K. Yang, A predictive analytics approach to reducing 30-day avoidable readmissions among patients with heart failure, acute myocardial infarction, pneumonia, or COPD. Health Care Manag. Sci. **18**(1), 19–34 (2015)
79. Y. Sun, B.H. Heng, S.Y. Tay, E. Seow, Predicting hospital admissions at emergency department triage using routine administrative data. Acad. Emerg. Med. **18**(8), 844–850 (2011)
80. K.B. Sundharakumar, S. Dhivya, S. Mohanavalli, R.V. Chander, Cloud based fuzzy healthcare system. Procedia Comput. Sci. **50**, 143–148 (2015)
81. A. Sunyaev, D. Chornyi, C. Mauro, H. Krcmar, Evaluation framework for personal health records: Microsoft HealthVault vs. Google Health, in *2010 43rd Hawaii International Conference on System Sciences*, (IEEE, Piscataway, 2010), pp. 1–10
82. R.K. Thomas, *Health Services Marketing: A Practitioner's Guide* (Springer Science & Business Media, New York, 2007)

83. J.E. Tym, C. Mitsopoulos, E.A. Coker, P. Razaz, A.C. Schierz, A.A. Antolin, B. Al-Lazikani, canSAR: An updated cancer research and drug discovery knowledgebase. Nucleic Acids Res. **44**(D1), D938–D943 (2016)
84. J. Vanschoren, J.N. van Rijn, B. Bischl, L. Torgo, OpenML: networked science in machine learning. ACM SIGKDD Explor. Newsl. **15**(2), 49–60 (2014)
85. D. Wang, E. Rundensteiner, R. Ellison, H. Wang, Probabilistic inference of object identifications for event stream analytics, in *Proceedings of the 16th International Conference on Extending Database Technology – EDBT '13*, (ACM Press, New York, 2013), p. 513
86. M.J. Ward, K.A. Marsolo, C.M. Froehle, Applications of business analytics in healthcare. Bus. Horiz. **57**(5), 571–582 (2014)
87. R.W. White, R. Harpaz, N.H. Shah, W. DuMouchel, E. Horvitz, Toward enhanced pharmacovigilance using patient-generated data on the internet. Clin. Pharmacol. Ther. **96**(2), 239–246 (2014)
88. R. Xu, Q. Wang, Large-scale extraction of accurate drug-disease treatment pairs from biomedical literature for drug repurposing. BMC Bioinformatics **14**, 181 (2013)
89. H. Yoshida, A. Kawaguchi, K. Tsuruya, Radial basis function-sparse partial least squares for application to brain imaging data. Comput. Math. Methods Med. **2013**, 591032 (2013)
90. X. Yu, A. Ganz, Scalable patients tracking framework for mass casualty incidents, in *Conference Proceedings: Annual International Conference of the IEEE Engineering in Medicine and Biology Society 2011*, (IEEE, Piscataway, 2011), pp. 860–863
91. Q. Zhang, C. Pang, S. Mcbride, D. Hansen, C. Cheung, M. Steyn, Towards health data stream analytics, in *IEEE/ICME International Conference on Complex Medical Engineering*, (IEEE, Piscataway, 2010), pp. 282–287
92. K. Zolfaghar, N. Verbiest, J. Agarwal, N. Meadem, S.-C. Chin, S.B. Roy, A. Teredesai, D. Hazel, P. Amoroso, L. Reed, Predicting risk-of-readmission for congestive heart failure patients: A multi-layer approach (2013). Available at: https://arxiv.org/abs/1306.2094